THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

The comparative proteomics of ubiquitination in mouse

Citation for published version:
RIKEN GER Group & Semple, CAM 2003, 'The comparative proteomics of ubiquitination in mouse', *Genome Research*, vol. 13, no. 6B, pp. 1389-94. https://doi.org/10.1101/gr.980303

Digital Object Identifier (DOI):
10.1101/gr.980303

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Genome Research

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 06. May. 2023

The Comparative Proteomics of Ubiquitination in Mouse

Colin A.M. Semple,[1,5] RIKEN GER Group[2] and GSL Members[3,4]

[1]MRC Human Genetics Unit, Crewe Road, Edinburgh, EH4 2XU, UK; [2]Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; [3]Genome Science Laboratory, RIKEN, Hirosawa, Wako, Saitama 351-0198, Japan

Ubiquitination is a common posttranslational modification in eukaryotic cells, influencing many fundamental cellular processes. Defects in ubiquitination and the processes it mediates are involved in many human disease states. The ubiquitination of a substrate involves four classes of enzymes: a ubiquitin-activating enzyme (E1), a ubiquitin-conjugating enzyme (E2), a ubiquitin protein ligase (E3), and a de-ubiquitinating enzyme (DUB). A substantial number of E1s (four), E2s (13), E3s (97), and DUBs (six) that were previously unknown in the mouse are included in the FANTOM2 Representative Transcript and Protein Set (RTPS). Many of the genes encoding these proteins will constitute promising candidates for involvement in disease. In addition, the RTPS provides the basis for the most comprehensive survey of ubiquitination-associated proteins across eukaryotes undertaken to date. Comparisons of these proteins across human and other organisms suggest that eukaryotic evolution has been associated with an increase in the number and diversity of E3s (possessing either zinc-finger RING, F-box, or HECT domains) and DUBs (containing the ubiquitin thiolesterase family 2 domain). These increases in numbers are too large to be accounted for by the presence of fragmentary proteins in the data sets examined. Much of this innovation appears to have been associated with the emergence of multicellular organisms, and subsequently of vertebrates, increasing the opportunity for complex regulation of ubiquitination-mediated cellular and developmental processes.

Ubiquitination is a central process in eukaryotic cells, influencing meiosis, cellular proliferation, and development. It is now seen as the most common posttranslational modification in all eukaryotes after phosphorylation. Until recently, it was thought that ubiquitin was simply a cellular label to target proteins for destruction in the proteasome, but it is now known to have additional roles in trafficking, kinase activation, and regulating gene expression (Conaway et al. 2002). These different roles are associated with different multi-ubiquitin chains, differing according to the particular lysine residue, within the ubiquitin molecules, which is used to interlink the chain (Weissman 2001). It appears that these different multi-ubiquitin chains, together with mono-ubiquitination, act as signals in different cell processes. Defects in ubiquitination are known to be involved in many human disease states such as developmental abnormalities, autoimmunity, neurodegenerative diseases (Alzheimer's disease, Parkinson's disease, Down's syndrome, and normal aging of the brain), and cancer (Weissman 2001). Ubiquitination is a multistep process requiring four classes of enzyme: a ubiquitin-activating enzyme (E1), a ubiquitin-conjugating enzyme (E2), a ubiquitin protein ligase (E3), and a de-ubiquitinating enzyme (DUB). Each of these four classes contain characteristic domains represented in the InterPro domain database (Apweiler et al. 2000).

Ubiquitination begins when the E1 recruits ubiquitin in an ATP-dependent process; then the E2 accepts ubiquitin from the E1. The E3 catalyzes the transfer of ubiquitin from the E2 to the substrate. At the proteasome, DUBs cleave multi-ubiquitin chains from substrates and disassemble the chains. This ensures that ubiquitinated proteins remain associated with the proteasome and prevents the accumulation of residual multi-ubiquitin chains that can disrupt proteasome activity. DUBs also process immature ubiquitin, which is translated as a fusion protein that is either a chain of ubiquitin molecules or joined to small ribosomal subunits (Weissman 2001). The flexibility and complexity of cellular regulation by ubiquitination can be increased at each stage by increasing the numbers of proteins in each of the four classes. For example, many E3s are SCF (Skp1-Cullin-F-box protein) complexes composed of proteins containing four InterPro domains: SKP1 component (IPR001232), F-box (IPR001810), Cullin (IPR001373), and zinc-finger RING (IPR001841; Jackson and Eldridge 2002). The large numbers of proteins containing these domains mean that a huge number of different E3 complexes, recognizing a wide range of protein substrates, can be formed.

The Representative Transcript and Protein Set (RTPS) amalgamates cDNA clones produced in the recent RIKEN FANTOM2 project with other publicly available cDNA sequences to provide the first overview of the mouse transcriptome (The FANTOM Consortium and the RIKEN GER Group Phase I and II Team 2002). Previous comparisons of ubiquitination-associated (UA) proteins between species such as human, *Caenorhabditis elegans*, and yeast have suggested that the numbers and compositions of E2 and E3 proteins are related to developmental complexity (Von Arnim 2001; Jones et al. 2002). The RTPS provides valuable additional data to investigate such assertions and also to seek putatively vertebrate specific features of ubiquitination. In the present study, we de-

[4]Takahiro Arakawa, Piero Caminci, Jun Kawai, and Yoshihide Hayashizaki.
[5]Corresponding author.
E-MAIL Colin.Semple@hgu.mrc.ac.uk; FAX 44 (0) 131-343-262.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.980303.

scribe the numbers and architectures of proteins containing UA InterPro domains across the known mouse proteome, in the light of the RTPS. These data are compared with the equivalent data from human, *Drosophila melanogaster*, *C. elegans*, *Schizosaccharomyces pombe*, and *Saccharomyces cerevisiae* proteomes. This study represents the most comprehensive survey of UA proteins across eukaryotes undertaken to date.

RESULTS

Table 1 shows the numbers of UA proteins present in three incomplete versions of the mouse proteome: the RTPS, a version derived from the SWISS-PROT and TrEMBL databases (SPTR), and a version derived from SWISS-PROT, TrEMBL, and the mouse Ensembl database (SPTRENS). It shows that the number of UA proteins within the RTPS exceeds the number within SPTR and constitutes more than three fourths of the number within SPTRENS. This suggests that the RTPS provides substantial coverage of the mouse transcriptome, because the SPTRENS includes proteins predicted from a high-quality mouse genome assembly produced by the Mouse Genome Sequencing Consortium (which has around seven times the coverage and covers an estimated 96% of the mouse euchromatic genome; see http://www.ensembl.org/Mus_musculus/). Of the 496 proteins present in the RTPS, 146 (30%) were found to be absent from the SPTRENS set (i.e., they lacked a matching sequence of ≥98% identity over ≥20 residues). Significant numbers of novel mouse proteins belonging to each enzyme class were discovered, though many have close human homologs.

The novel UA proteins from the RTPS were combined with those from the SPTRENS set to provide the most complete set of mouse UA proteins possible; this combined set was then compared with sets of UA proteins from five other "draft" or completely sequenced organisms. Table 2 shows that higher numbers of E1 and E2 proteins are found in mouse and human than in the other organisms examined. This is also true of E3 proteins if one ignores the very high number of F-box (IPR001810) proteins in *C. elegans* (which remains mysterious; Kipreos and Pagano 2000), and the same observations can also be made of the remaining two classes (DUB and nonapplicable [N/A] proteins). The differences and similarities between these organisms are often clearer at the level of the domains that define the enzyme classes. Within E3 proteins, the largest differences between organisms are seen in zinc-finger RING (IPR001841) or HECT (IPR000569) domains in particular, whereas within DUBs, the ubiquitin thiolesterase family 2 (IPR001394) domain shows the largest differences. Within the N/A class, the largest differences are seen for the ubiquitin interacting motif (UIM; IPR003903) and the UBA domain (IPR000626), but also for the UBX domain (IPR001012) and the CUE domain (IPR003892).

Comparisons between species that rely solely upon the numbers of proteins within each enzyme class omit valuable information about the architectures of the proteins concerned. In particular, many of the UA proteins identified possess additional InterPro domains, not directly related to ubiquitination, with known or inferred functions ("non-UA domains"). Table 3 shows the numbers of additional, non-UA domains in UA proteins from each class. From Table 3, it is possible to see the ratio of the numbers of non-UA domains to the number of proteins they are present within. Usually, this ratio is ~1, reflecting little redundancy in domain composi-

Table 1. Mouse Ubiquitination-Associated Proteins From Three Different Data Sets, Categorized by InterPro Domains and Enzyme Class

Class	InterPro acc	RTPS	SPTR	SPTRENS	InterPro domain name
E1	IPR000594	12 (4)	8	9	UBA/THIF-type NAD/FAD binding fold
	IPR000127	6 (1)	6	7	Ubiquitin-activating enzyme repeat domain
	IPR000011	2 (0)	4	4	Ubiquitin-activating enzyme
Total E1		12 (4)	14	16	
E2	IPR000608	40 (13)	24	48	Ubiquitin-conjugating enzymes
E3	IPR001810	56 (21)	34	53	F-box
	IPR001373	9 (5)	2	4	Cullin
	IPR000569	24 (5)	14	33	HECT domain (ubiquitin-protein ligase)
	IPR003126	3 (1)	1	7	Zinc-finger in N-recognin (putative)
	IPR001841	192 (61)	143	244	Zinc-finger, RING
	IPR001232	2 (0)	2	3	SKP1 component
	IPR003613	8 (4)	3	4	Zinc-finger, modified RING (U-box)
Total E3		292 (97)	287	431	
DUB	IPR001578	5 (0)	5	5	Ubiquitin C-terminal hydrolase, family 1
	IPR001394	40 (5)	30	65	Ubiquitin thiolesterase, family 2
	IPR001607	10 (2)	8	19	Zinc-finger in ubiquitin thiolesterase
Total DUB		48 (6)	44	79	
N/A	IPR000626	48 (12)	38	70	Ubiquitin domain
	IPR000449	38 (9)	32	60	Ubiquitin-associated (UBA) domain
	IPR001012	11 (6)	6	11	UBX domain
	IPR003892	8 (1)	6	10	CUE domain
	IPR003903	15 (3)	12	16	Ubiquitin interacting motif (UIM)
	IPR004854	1 (0)	1	1	Ubiquitin fusion degradation protein UFD1
Total N/A		101 (26)	103	173	
	TOTAL	496 (146)	352	630	

The numbers of novel proteins, not present in the other sets, are indicated in parentheses. The N/A class contains proteins possessing domains that do not yet have well understood roles in the other four classes.
RTPS indicates Representative Transcript and Protein Set; SPTR, SWISSPROT and TREMBL databases; and SPTRENS, SWISSPROT, TREMBLE, and mouse Ensembl databases.

Table 2. The Numbers of Ubiquitination-Associated Proteins Encoded in Six Draft/Finished Eukaryotic Genomes, Categorized by InterPro Domains and Enzyme Class

Class	InterPro acc	*Homo sapiens*	*Mus musculus*	*Drosophila melanogaster*	*Caenorhabditis elegans*	*Schizosaccharomyces pombe*	*Saccharomyces cerevisiae*
E1	IPR000594	14	13	11	9	9	8
	IPR000127	7	8	3	4	3	3
	IPR000011	2	4	1	1	2	2
Total E1		16	16	11	9	9	8
E2	IPR000608	53	61	32	25	14	15
E3	IPR001810	78	74	34	434	13	14
	IPR001373	10	9	7	10	4	5
	IPR000569	38	38	13	10	7	5
	IPR003126	7	8	6	6	3	2
	IPR001841	385	305	121	170	51	40
	IPR001232	6	3	8	26	2	2
	IPR003613	9	8	6	4	3	2
Total E3		527	442	189	657	82	68
DUB	IPR001578	4	5	5	3	2	1
	IPR001394	67	70	21	33	16	18
	IPR001607	14	21	5	6	6	4
Total DUB		74	78	27	39	20	20
N/A	IPR000626	69	82	27	30	17	9
	IPR000449	65	69	28	19	12	9
	IPR001012	19	17	8	5	5	7
	IPR003892	11	11	4	4	3	7
	IPR003903	30	19	8	7	6	7
	IPR004854	1	1	1	2	1	1
Total N/A		184	187	63	57	39	35
TOTAL		835	764	320	785	162	145

The N/A class contains proteins possessing domains that do not yet have well understood roles in the other four classes.

tion between proteins within a class; this trend is clearest within E1 and E2 proteins. The most noticeable exceptions to this are human and *C. elegans* E3 proteins, in which the various non-UA domains are distributed among more than 2.5 times the number of E3 proteins. More modest levels of redundancy in domain composition are seen within E3 proteins from every other organism except *S. cerevisiae*, within DUB proteins from most organisms and within human and mouse N/A proteins. The full domain architecture for any protein encoded by a FANTOM2 clone can be examined in the online FANTOMDB annotation database (http://fantom2.gsc.riken.go.jp/db/), which can be searched with InterPro domain accession numbers.

Eight non-UA domains were found to be present within UA proteins from every species examined: NAD binding site (IPR000205), RNA-binding region RNP-1 (IPR000504; found in a variety of RNA binding proteins), sevenfold repeat in clathrin and VPS proteins (IPR000547; may be involved with endocytosis), zinc-finger, C-x8-C-x5-C-x3-H type (IPR000571; involved in cell cycle– or growth phase–related regulation and splicing), zinc-finger C2H2 type (IPR000822; responsible for RNA and DNA binding), TPR repeat (IPR001440; a protein–protein interaction motif), G protein-β WD-40 repeat (IPR001680; present in signal transduction proteins), and zinc-finger cysteine-rich C6HC (IPR002867; function unknown). All of these domains are found in combination with E3 proteins containing either zinc-finger RING (IPR001841) or F-box (IPR001810) domains within every species. In addition, within the human proteome, two of these non-UA domains (IPR000822, IPR001440) have been recruited to ubiquitin thiolesterase family 2 (IPR001394) containing DUBs. A further four non-UA domains were found within UA proteins from every multicellular organism studied but at the same time were absent from UA proteins from the two yeast species: zinc-finger ZZ type (IPR000433), basic helix-loop-helix dimerization domain (IPR001092), tyrosine protein kinase

Table 3. The Numbers of Domains Not Directly Related to Ubiqitination Within Ubiquitination-Associated Proteins From Each Class

Class	*Homo sapiens*	*Mus musculus*	*Drosophila melanogaster*	*Caenorhabditis elegans*	*Schizosaccharomyces pombe*	*Saccharomyces cerevisiae*
E1	4 (9)	7 (16)	8 (9)	5 (7)	4 (6)	2 (4)
E2	5 (4)	1 (1)	7 (7)	3 (4)	0 (0)	0 (0)
E3	114 (325)	49 (62)	87 (101)	123 (321)	46 (49)	32 (30)
DUB	17 (21)	4 (9)	11 (8)	14 (18)	10 (9)	6 (8)
N/A	61 (99)	18 (29)	49 (39)	41 (31)	36 (24)	20 (16)
Total	147 (430)	57 (64)	128 (151)	156 (365)	79 (80)	47 (50)

Numbers of proteins in parentheses.

(IPR001245), and ankyrin repeat (IPR002110). Apart from the tyrosine protein kinase domain (IPR001245), these non-UA domains are found within E3 proteins, in combination with either zinc-finger RING (IPR001841) or HECT (IPR000569) domains. The Basic helix-loop-helix dimerization domain (IPR001092) is also found in a human ubiquitin thiolesterase family 2 (IPR001394) containing DUB. The tyrosine protein kinase domain (IPR001245), on the other hand, is found in proteins from the N/A class that contain the UBA domain (IPR000449); there are eight human proteins carrying this combination, one in mouse, two in *D. melanogaster*, and one in *C. elegans*.

Only four non-UA domains were found to be specific to mouse and human UA proteins: HMG-I and HMG-Y DNA-binding domain (IPR000637), sterile α motif SAM (IPR001660), butyrophylin-like (IPR003879), and fibronectin type III repeat (IPR003962). All four of these domains appear in combination within the zinc-finger RING domain (IPR001841). Both the butyrophylin-like (IPR003879) domain and the fibronectin type III repeat (IPR003962) are absent within zinc-finger RING (IPR001841) proteins from the Ensembl Fugu Database.

DISCUSSION

This analysis suggests that, as stated previously (Von Armin 2001; Jones et al. 2002), there may be a direct relationship between the number of E2 proteins and developmental complexity, but also that a similar case can be made for the numbers of proteins occupying the other enzyme classes. It is clear that most evolutionary innovation in UA proteins has occurred within E3 proteins, particularly those containing zinc-finger RING (IPR001841), F-box (IPR001810), or HECT (IPR000569) domains. However, lineage-specific explosions in E3 numbers, as with the F-box (IPR001810)—containing proteins in *C. elegans*, may obscure the real relationship between the numbers of proteins in an enzyme class and another parameter such as developmental complexity. In addition, because E3 proteins are concerned with substrate recognition, one would expect to see their numbers directly related to total proteome size, which should confound any relationship with developmental complexity. It seems premature to draw any firm conclusions about ubiquitination and development from such comparisons. It may be more informative to examine differences between organisms at the level of the component domains for each class, both in terms of the numbers and architectures of the proteins that possess these domains.

The differences in numbers of ubiquitin thiolesterase family 2 DUBs and proteins containing the N/A class domains, such as UIM (IPR003903) and UBA (IPR000449), are arguably the most intriguing identified here. Early vertebrate evolution seems to have involved a substantial increase in ubiquitin thiolesterase family 2 domain (IPR001394) DUBs, because there are 104 in the Ensembl Fugu database (an unknown number of these may be fragmentary predictions and/or encoded by pseudogenes). Most DUBs seem to have roles in reversing, proof-reading, and introducing variations in ubiquitination (Ben-Neriah 2002). The data here suggest that as well as ubiquitin conjugation and ligation, the de-ubiquitination process has also been exploited to achieve ever more intricate regulation of proteolysis in vertebrates. The roles of several domains examined here have yet to be fully described, although recently some progress has been made with the UIM (IPR003903). The UIM has been shown to be necessary for the mono-ubiquitination of mouse protein substrates, leading to their inclusion in endocytic vesicles (Polo et al. 2002). The substrates then become cargo in intracellular membrane trafficking pathways and an extensive UIM-ubiquitin-based intracellular network has been postulated on the basis of these results (Polo et al. 2002). In this context, it is interesting that vertebrate evolution has involved an increase in UIM-containing proteins relative to other sequenced organisms, presumably increasing the size of any associated intracellular network. The Ensembl Fugu database contains only up to 13 UIM-containing proteins, suggesting that the elaboration of this network has continued since the divergence of fish and mammals. The UBA domain (IPR000449) is likely to be a general multi-ubiquitin binding domain (Wilkinson et al. 2001) but also seems to mediate dimerization (Bertolaet et al. 2001). The Ensembl Fugu database contains up to 79 UBA containing proteins, and so, this domain also seems to have expanded in numbers since the emergence of vertebrates, although perhaps not since the divergence of fish and mammals.

If one wishes to draw conclusions about the different cellular and developmental roles that ubiquitination has evolved to regulate, then it is instructive to study the broader architectural differences between UA classes across different organisms. Since the emergence of multicellular eukaryotes, three domains seem to have been added to the repertoire of E3 (and human DUB) proteins: the zinc-finger ZZ type domain (which has an unknown function), the basic helix-loop-helix dimerization domain (which is present in transcription factors, influencing a variety of developmental pathways), and the ankyrin repeat (which functions as a protein–protein interaction domain). During the same time, the tyrosine protein kinase domain (which is involved in the response of eukaryotic cells to external stimuli) has been partnered with the UBA domain. Since the evolution of vertebrates, the HMG-I and HMG-Y DNA-binding domain (which may function in mRNA processing) and sterile α motif (which is believed to be involved in the regulation of numerous developmental processes) have been added to zinc-finger RING domain E3 proteins. At a later stage in vertebrate evolution, after the divergence of fish and mammals, this same family of E3 proteins has acquired the butyrophylin-like domain (which has no known function but is found within transcription factors, ribonucleoproteins, and proto-oncoproteins) and the fibronectin type III repeat (which is found in a variety of cell surface binding proteins).

Future work uncovering the cellular roles of ubiquitination that are universal across eukaryotes may benefit from investigation of the non-UA domains found to be present within E3 proteins from every species examined. The functions of these non-UA domains cover signaling, endocytosis, splicing, and transcriptional regulation. Some fundamental observations can be made about the progressive integration of ubiquitination into the life of the eukaryotic cell throughout evolution. Several cellular processes have been linked to ubiquitination since the emergence of eukaryotes: the cell cycle, RNA processing and the regulation of transcription, intracellular transport, and signaling pathways. Throughout eukaryotic evolution, an ever more diverse variety of links have appeared between ubiquitination and such cellular processes. These links have predominantly been achieved through innovations in the architecture of three families of E3s (possessing zinc-finger RING, F-box, or HECT domains) and the ubiq-

uitin thiolesterase family 2 DUBs. However, other, presently less well understood families of UA proteins (possessing UBA, UIM, and other domains) have also contributed to the diversity of these links. It would appear that the emergence of multicellular organisms and then of vertebrates have been associated with much of this innovation, which has led to novel connections with RNA processing and signaling as well as involvement in the regulation of developmental pathways. This steady increase in the cellular connectivity of ubiquitination is likely to have increased the opportunity for fine regulation of the associated cellular processes.

METHODS

A nonredundant, although incomplete, *Mus musculus* proteome (based upon proteins from SWISS-PROT and TrEMBL and additional peptides predicted by Ensembl) was obtained from the EBI Proteome Analysis Database (http://www.ebi.ac.uk/proteome/; Apweiler et al. 2001). The EBI data set retrieved (30,900 proteins as of August 28, 2002) was altered to remove all proteins predicted from FANTOM2 clone sequences, as indicated in either the name or description fields of the protein entry. This altered EBI data set appears as SPTRENS in Table 1. This alteration allowed the real contribution of novel proteins from RTPS6 to be calculated. UA proteins from both the SPTRENS set and the RTPS6 were defined by the presence of the InterPro domains listed below. This SPTRENS set was then combined with proteins from the RTPS6 to produce the total UA *M. musculus* proteins referred to in Table 2. Matches between RTPS6 proteins and those from SPTRENS of ≥98% identity over ≥20 residues were assumed to indicate that both sequences originated from the same gene. This may underestimate the actual number of genes if the data set contains many recent paralogs. An additional *M. musculus* proteome, composed only of proteins from SWISS-PROT and TrEMBL (SPTR in Table 1), was also retrieved from the EBI Proteome Analysis Database, to estimate the relative success of novel gene identification strategies based upon genomic or transcribed sequence. At the same time (August, 28, 2002), nonredundant protein sets for *Homo sapiens* (33,451 protein entries derived from SWISS-PROT, TrEMBL and Ensembl), *C. elegans* (20,046 proteins), *D. melanogaster* (15,253 proteins), *S. pombe* (5051), and *S. cerevisiae* (6206 proteins) proteomes were also retrieved from the same source. The methods used to reduce redundancy vary between the proteome sets referred to here. The FANTOM2 RTPS6 set was constructed by using mapping to the genome to reduce redundancy, and when a finished genome sequence is available, the EBI follows an analogous strategy. However, when a finished genome sequence is not available (for mouse and human sets), mapping to the genome sequences cannot always provide a definitive answer. Consequently, the EBI uses another method of reducing redundancy: liberal clustering of all proteins sharing sequence similarity, with each cluster treated as a single database entry. These strategies are discussed in detail at the EBI Proteome Analysis Database site (http://www.ebi.ac.uk/proteome/). All data sets used in this study are available from http://www.hgu.mrc.ac.uk/Users/Colin.Semple/lab_data.html.

Given the incomplete nature of the sequence data for some of these organisms and the degree of error inherent in gene prediction, each of these data sets would be expected to contain a number of incomplete fragmentary protein sequences. It is not a trivial task to derive accurate estimates of these numbers. However, to assess the scale of this problem, the proportions of proteins from each proteome indicated to be fragmentary (within the database name or description fields) were recorded as follows: *H. sapiens*, 0.12; *M. musculus*, 0.09; *D. melanogster*, 0.01; *C. elegans*, 0.001; *S. pombe*, 0.05; and *S. cerevisiae*, 0.01. This indicates that fragmentary proteins are most prevalent in *H. sapiens* and *M. musculus* data, and the equivalent proportions for UA proteins in these species are 0.13 and 0.05, respectively. One may conclude that some caution should be exercised in interpreting comparisons involving these two proteomes. Some comparisons involved predicted proteins from the recently sequenced draft genome sequence of the puffer fish *Fugu rubripes* (Ensembl Fugu Database Release 8.1.1; http://www.ensembl.org/Fugu_rubripes/). Given the lack of transcribed sequence data for *F. rubripes* and the unfinished state of its genome sequence, these predicted proteins should also be regarded as potentially incomplete, fragmentary and/or artifactual.

The UA InterPro domains examined were as follows:

1. the three E1 domains: UBA/THIF-type NAD/FAD binding fold domain (IPR000594), E1 replicated domain (IPR000127), and E1 domain (IPR000011);
2. the single E2 domain: E2 domain (IPR000608);
3. the seven E3 domains: SKP1 component domain (IPR001232), F-box domain (IPR001810), Cullin domain (IPR001373), HECT domain (IPR000569), zinc-finger (putative), N-recognin domain (IPR003126), zinc-finger RING domain (IPR001841), and zinc-finger–modified RING domain (also called Ubox; IPR003613);
4. the three DUB domains: ubiquitin C-terminal hydrolase family 1 domain (IPR001578), ubiquitin thiolesterase family 2 domain (IPR001394), zinc-finger in ubiquitin thiolesterase domain (IPR001607);
5. six other domains (assigned to the N/A class because other classes are not applicable at present), for which we lack a precise knowledge of their roles in ubiquitination, were also included: ubiquitin domain (IPR000626; a number of proteins contain ubiquitin-like domains with unknown functions that are similar to the ubiquitin molecule itself), UBA domain (IPR000449; recently shown to bind multiubiquitin chains), UBX domain (IPR001012; function unknown but may be involved in apoptosis), CUE domain (IPR003892; may be involved in binding E2s), ubiquitin interacting motif (IPR003903; involved with monoubiquitination of substrates leading to endocytosis), and ubiquitin fusion degradation protein UFD1 (IPR004854; involved in the recognition of ubiquitin–protein conjugates leading to degradation).

Unless otherwise stated, the functional description of these domains is derived from the appropriate InterPro entries. Contemporary references for each domain can also be viewed at the InterPro site (http://www.ebi.ac.uk/interpro/index.html). It is important to note that InterPro assignments are not infallible, and occasionally, a single domain may be annotated as belonging to two different categories; an example encountered in these data was completely overlapping predictions of RING finger (IPR001841) and PHD finger (IPR001965) domains. Annotation for all proteins of interest can be viewed in its entirety using the graphical annotation viewer provided by FANTOMDB (http://fantom2.gsc.riken.go.jp/db/).

ACKNOWLEDGMENTS

This work was supported by the U.K. Medical Research Council. It also benefited from many discussions with other FANTOM Consortium members and the hospitality of the RIKEN Genomic Sciences Center. Colin Gordon read an earlier version of the manuscript, and an anonymous reviewer provided useful comments.

REFERENCES

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. 2000. InterPro: An integrated documentation resource for

protein families, domains and functional sites. *Bioinformatics* **16:** 1145–1150.
Apweiler, R., Biswas, M., Fleischmann, W., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E.V., Mittard, V., Mulder, N., Phan, I., et al. 2001. Proteome Analysis Database: Online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Res.* **29:** 44–48.
Ben-Neriah, Y. 2002. Regulatory functions of ubiquitination in the immune system. *Nat. Immunol.* **3:** 20–26.
Bertolaet, B.L., Clarke, D.J., Wolff. M., Watson, M.H., Henze, M., Divita, G., and Reed, S.I. 2001. UBA domains mediate protein–protein interactions between two DNA damage-inducible proteins. *J. Mol. Biol.* **313:** 955–963.
Conaway, R.C., Brower, C.S., and Conaway, J.W. 2002. Emerging roles of ubiquitin in transcription regulation. *Science* **296:** 1254–1258.
The FANTOM Consortium and the RIKEN GER Group Phase I and Phase II Team. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.
Jackson, P.K. and Eldridge A.G. 2002. The SCF ubiquitin ligase: An extended look. *Mol. Cell* **9:** 923–925.
Jones, D., Crowe, E., Stevens, T.A., and Candido, E.P. 2002. Functional and phylogenetic analysis of the ubiquitylation system in *Caenorhabditis elegans*: Ubiquitin-conjugating enzymes, ubiquitin-activating enzymes, and ubiquitin-like proteins. *Genome Biol.* **3:** RESEARCH0002.
Kipreos, E.T. and Pagano, M. 2000. The F-box protein family. *Genome Biol.* **1:** REVIEWS3002.
Polo, S., Sigismund, S., Faretta, M., Guidi, M., Capua, M.R., Bossi, G., Chen, H., De Camilli, P., and Di Fiore, P.P. 2002. A single motif responsible for ubiquitin recognition and monoubiquitination in endocytic proteins. *Nature* **416:** 451–455.
Von Arnim, A.G. 2001. A hitchhiker's guide to the proteasome. SciSTKE 2001(97):PE2.
Weissman, A.M. 2001. Themes and variations on ubiquitylation. *Nat. Rev. Mol. Cell. Biol.* **2:** 169–178.
Wilkinson, C.R.M., Seeger, M., Hartmann-Petersen, R., Stone, M., Wallace, M., Semple, C., and Gordon, C. 2001. Proteins containing the UBA domain are able to bind to multi-ubiquitin chains. *Nat. Cell. Biol.* **3:** 939–943.

WEB SITE REFERENCES

http://www.ensembl.org/Mus_musculus/; Mouse Genome Sequencing Consortium.
http://fantom2.gsc.riken.go.jp/db/; FANTOMDB annotation database.
http://www.ebi.ac.uk/proteome; EBI Proteome Analysis Database.
http://www.hgu.mrc.ac.uk/Users/Colin.Semple/lab_data.html; all data sets used in this study.
http://www.ensembl.org/Fugu_rubripes/; draft genome sequence of the puffer fish.
http://www.ebi.ac.uk/interpro/index.html; InterPro site.

Received November 8, 2002; accepted in revised form March 6, 2003.